Engenheiro(a) de dados

- Gostaríamos de analisar suas habilidades com SQL, modelagem dimensional e integração de dados. Mostre seus conhecimento em processos de ETL e conceitos de Data Warehouse? Que tal replicar nossos datasets, remodelar em um banco de dados e apresentar as melhorias realizadas em sua criação?

  Anexo.

- É possível utilizar o modelo proposto em um ambiente cloud? Quais plataformas ou serviços você utilizaria? Quais as vantagens do modelo escolhido em questões de performance?

  Resposta: Sim, é possível. Utilizaria a plataforma da AWS que atualmente é referência em computação sob demanda, possibilitando escalar o ambiente em paralelo ao crescimento dos dados coletados e armazenados. O modelo escolhido privilegia o acesso aos dados sem a necessidade de realização de *'joins'* para obtenção dos dados, com a criação de índices adequados e a não concorrência com o ambiente OLTP, ganharemos em performance.

- Alguns membros do time dizem que a atual modelagem do banco de dados é adequada para o uso dos cientistas de dados e analistas de BI, porém, outros dizem que existem formas de modelar bancos de dados que trarão mais eficiência. Qual é a sua opinião sobre isso?

  Resposta: A modelagem atual pode atender sim, porém os dados estão normalizados na Segunda Forma Normal e isso implica em construir consultas que façam união entre várias tabelas para se obter a informação necessária. Além disso, para utilização de métricas, precisaríamos realizar agrupamentos nas consultas (GROUP BY) para construir as agregações, o que também gera perda de performance. Quando trabalhamos com modelos OLAP, criamos dimensões e tabelas fato que conseguem fazer agregações sem a necessidade de realizar os agrupamentos em instruções SQL. Sendo assim, existem formas que trarão mais eficiência, seguindo os conceitos de Data Warehouse.

- Estamos preocupados com o vertiginoso aumento do volume em nosso banco de dados atual? Você consideraria uma opção mais escalável ou devemos manter a estrutura existente?

  Resposta: No exercício não ficou claro como é a estrutura atual do banco de dados em termos de recursos computacionais, porém, considerando a introdução que informa sobre 750 mil produtos, centenas de milhares de pedidos e mais de 5 mil lojistas, torna-se imprescindível investir em computação em nuvem, uma opção mais segura e que permite trabalhar com recursos sob demanda, priorizando também controle de custos operacionais.

- Nossa ferramenta de visualização de dashboards está lenta e o nosso time detectou que o problema está na infraestrutura de dados. Como você abordaria esta situação do ponto vista de arquitetura de dados?

  Resposta: O primeiro passo é entender como funciona o fluxo de dados, a fim de identificar se há fragmentação. Analisar o modelo OLAP revisando os relacionamentos construídos dentro dos cubos, observar se há dados sendo exibidos sem frequência de utilização e a partir destes pontos adotar ações de correção.

- Nosso banco de dados está hospedado na nuvem e nossas ferramentas de análise de dados são "on premisses". Você manteria este arranjo ou faria mudanças visando mais performance?

  Resposta: Após analisar as variáveis de custo, desempenho e disponibilidade de recursos para mudanças, indicaria que banco de dados e ferramentas de análise de dados fizessem parte da mesma rede a fim de evitar possíveis gargalos por conta de comunicação.

- Nossa área operacional necessita de informações em tempo real, porém os diretores da empresa, que acompanham somente informações de KPIs mensais, alegam que isso é desnecessário e acarretaria custos. Qual é o seu posicionamento sobre isso?

  Resposta: Caso a análise de custos realmente prove que não é viável a criação e disponibilização de visões para a área operacional, pode-se adotar rotinas que seriam executadas periodicamente que façam a leitura do DW e disponibilize esses dados via e-mail ou em arquivos em diretórios compartilhados, por exemplo.

- Nosso time que está focado em Governança de Dados alega que documentar os processos é mais importante do que refatorar os mais de 500 scripts que estão funcionando com lentidão. Como você atuaria neste impasse, se tivesse que priorizar o trabalho?

  Resposta: A lentidão apresentada pelos scripts desenvolvidos anteriormente pode gerar impacto na operação e afetar os clientes, desta forma, priorizaria a refatoração dos scripts, documentado-os simultaneamente. Por fim, caso existam scripts dentre os mais de 500 que não seja necessário realizar refatoração, estes seriam somente documentados.

- Aqui no olist, somos muito mão na massa! Como Engenheiro(a) de dados, mostre pra gente o que você consegue fazer na prática com esse nosso banco de dados. (Sabemos que é uma amostra, mas imagine que o todo pode ser petabytes de dados)

  Anexo.

- O que acha de escrever um relatório ou slides sobre a sua abordagem na solução de alguns desses problemas?

  Anexo.

- Fique livre para criar sua própria abordagem, caso considere que as dicas anteriores não sejam pertinentes.

ANEXO

Abordagem

Considero a estrutura transacional eficaz para consulta, apenas realizando a criação de índices, porém há organizações que podem trazer otimização nas consultas, tais como:

- Modelagem em dimensões e fatos (métricas)
- União de dados pertinentes em uma mesma tabela a fim de evitar 'joins' que podem prejudicar a performance
- Importação somente de dados que são relevantes para os 'insights' da organização, desconsiderando dados de controle do transacional

Com acesso a uma quantidade maior de dados, creio que poderia desenvolver estruturas mais completas relacionadas entre si.

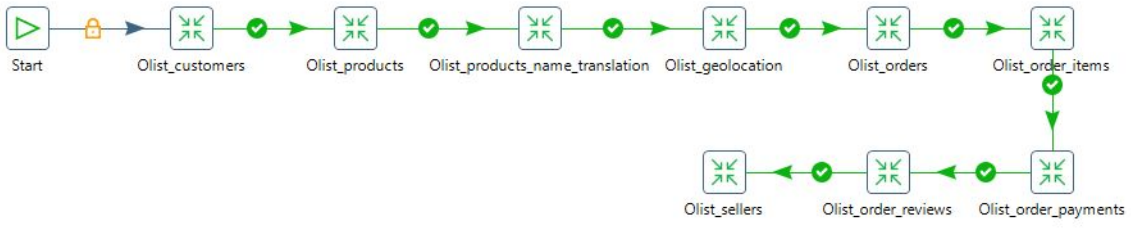
Ferramentas

Para desenvolver este trabalho, utilizei as seguintes ferramentas:

- Pentaho Data Integration (Kettle): para ETL;
- MySQL como SGBD;
- SQL Power Architect para modelagem do DW;
- Metabase;
- Desenvolvi meu trabalho localmente, mas para aplicações corporativas, concordo que o desenvolvimento é mais adequado em ambiente cloud

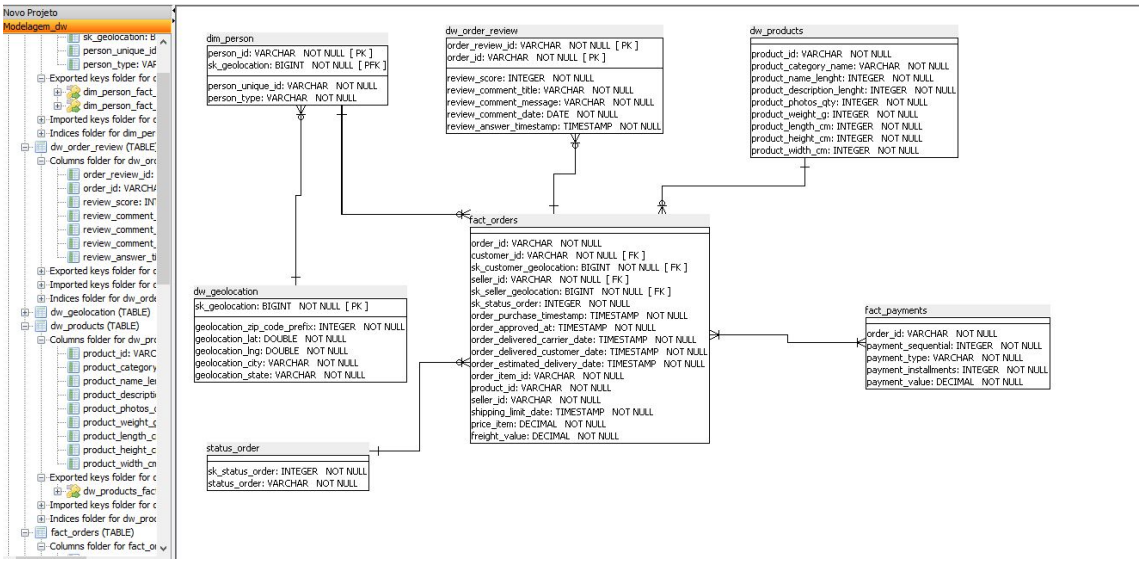
Desenvolvimento

Iniciei meu trabalho realizando a importação dos datasets através de um job criado no PDI (Pentaho Data Integration):

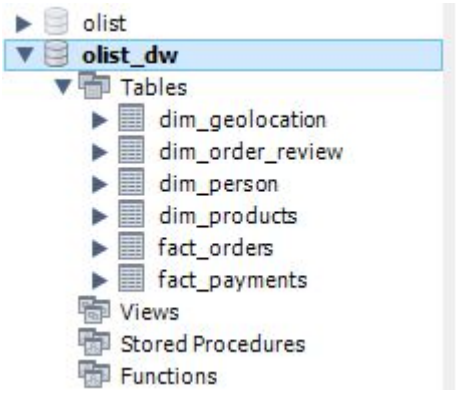


O job acima é formado por transformações que basicamente fazem a leitura do arquivo .csv e em seguida, insere no banco de dados MySQL:

Finalizada a inserção das informações contidas nos datasets, fiz o esboço do DW no SQL Power Architect:



A partir daí, executei a Engenharia Reversa da ferramenta para criar o script DDL. Entretanto, identifiquei algumas melhorias e as realizei no próprio script ao executá-lo no MySQL.



Para o modelo acima, adotei os seguintes critérios:

- Criei a dimensão 'person' para conter todas as pessoas: 'customers' e 'sellers', sendo diferenciadas pelo seu tipo: 'C' ou 'S'.

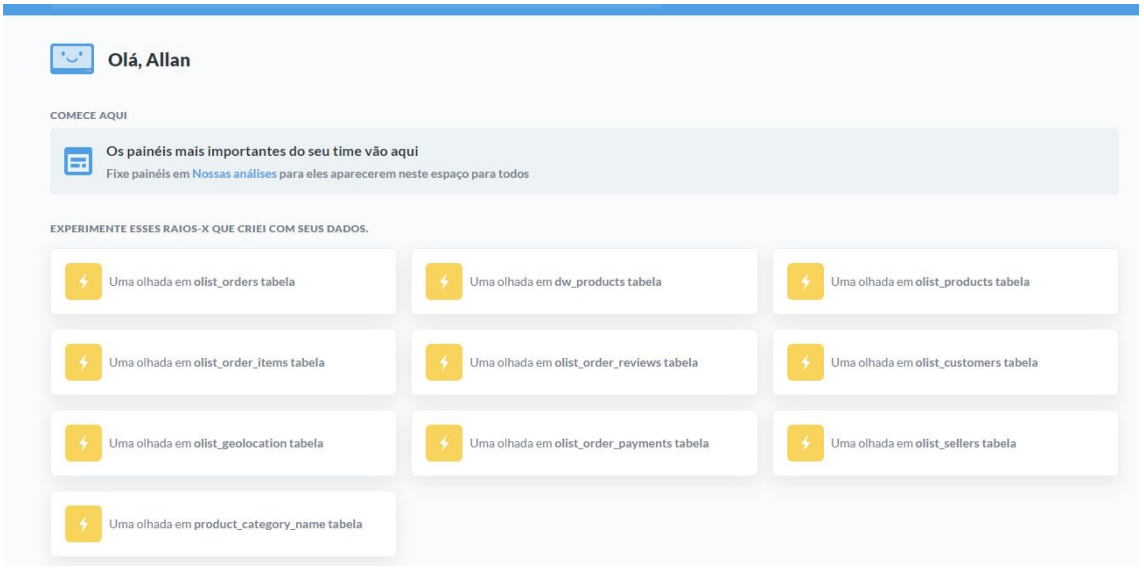
- Criei a fato ‘orders’ que contempla os dados do cabeçalho, produtos, clientes, vendedores e itens, ligados através de suas respectivas chaves, com otimização através de índices.
- Para as demais tabelas, mantive a mesma estrutura, pois creio que estão adequadas.

Utilizei novamente o PDI (Pentaho Data Integration) para realizar a carga de dados para o modelo dw:

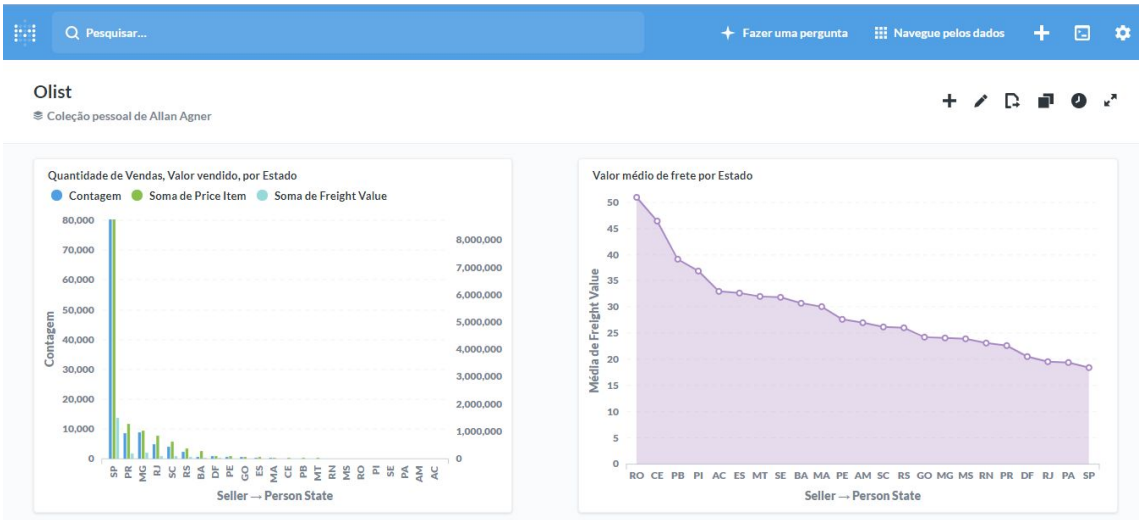
Visualização

Para demonstrar a praticidade de construção de visões a partir dos dados importados, construí o exemplo abaixo, utilizando o Metabase, uma ferramenta open source para criação de análises de dados.

Página inicial com a estrutura de dados:



Exemplos de gráficos utilizando os dados:



Conclusão

Com a organização dos dados sugerida acima, creio que haverá um melhor entendimento da equipe de Ciência de dados e BI, que não terão dificuldade em abstrair suas visões a partir dos dados armazenados, pois estão armazenados de forma intuitiva e de fácil utilização para visões analíticas ou sintéticas.